SYNOPSIS
of the Ph. D Thesis titled

DEVELOPMENT OF SPEECH EMOTION RECOGNITION USING DEEP LEARNING

SUBMITTED TO

Faculty of Basic and Applied Science,
MGM University
Chhatrapati Sambhajinagar (MS) - 431003, INDIA

by

IRFAN ABDULLATIF CHAUGULE

in partial fulfilment for the requirements of the award of the degree of

DOCTOR OF PHILOSOPHY
IN COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

UNDER THE SUPERVISION OF

DR. SATISH R. SANKAYE
Associate Professor, MGM University, DR.G.Y. Pathrikar College of Computer Science and Information Technology, Chhatrapati Sambhajinagar, Maharashtra, India

(JUN 2025)

TABLE OF CONTENTS

1 Introduction to Affective Computing and the Speech Emotion Recognition Challenge

1.1 The Role of Emotion in Human-Computer Interaction

1.2 The Evolution of Automated Emotion Recognition: From Handcrafted Features to Deep Learning Paradigms

1.3 Core Challenges in Speech-Based Emotion Analysis

1.4 Problem Statement and Research Objectives

2 A Critical Review of the State-of-the-Art in Speech Emotion Recognition

2.1 Paradigms in Feature Extraction: A Comparative Analysis

2.2 Traditional Machine Learning Approaches: Strengths and Limitations

2.3 The Deep Learning Revolution in SER

2.4 Identifying the Research Gap

3 Foundational Methodologies and Architectural Components

3.1 Pilot Study: Existing Work on Classical Machine Learning Models (RAVDESS)

3.2 Why Deep Learning Models Are Required and Superior

3.3 Enhancing Data Diversity: A Deep Dive into Augmentation Techniques for Speech Signals

3.4 Time-Frequency Analysis: The Stockwell Transform and its Boosted Variant

3.5 Model Compression and Knowledge Transfer: The Principles of Knowledge Distillation

3.6 Architectural Deep Dive: The Teacher Network (RPAE-BiGRU)

3.7 Architectural Deep Dive: The Student Network (CCN)

4 The Proposed Hierarchical SER Framework

4.1 System Overview and End-to-End Workflow

4.2 Corpus Analysis and Preparation 4.3 The Signal Processing Pipeline: From Raw Audio to Spectrogram Representation

4.3 Detailed Architecture of the RPAE-BiGRU Teacher Model

4.4 Detailed Architecture of the Convolutional Capsule Network Student Model

4.5 The Knowledge Distillation Protocol

5 Empirical Evaluation and Performance Analysis

5.1 Experimental Design, Hyperparameter Tuning, and Implementation Details
5.2 Performance Analysis on Individual Datasets
5.3 Comprehensive Comparative Analysis
5.4 Ablation Studies: Quantifying the Contribution of Individual Components
6 Discussion, Conclusion, and Future Trajectories
6.1 Synthesis of Findings and Interpretation of Results
6.2 Primary Contributions to the Field of Speech Emotion Recognition
6.3 Limitations of the Proposed Framework
6.4 Future Research Directions and Potential Applications
7 References

1: Introduction to Affective Computing and the Speech Emotion Recognition Challenge

1.1 The Role of Emotion in Human-Computer Interaction

Human communication is a multifaceted process that extends far beyond the literal meaning of words. Speech, as the most natural and pervasive mode of interaction, serves as a rich conduit for both linguistic and paralinguistic information. Among the most significant forms of paralinguistic information is emotion, a fundamental aspect of the human experience that provides a window into an individual's mental and physical state, intentions, and opinions. The ability to perceive and interpret these emotional cues is critical for effective and empathetic communication (Gunawan et al., 2021; Akçay & Oğuz, 2020; Alhussain et al., 2019).

In the rapidly advancing field of Human-Computer Interaction (HCI), the pursuit of more natural and intuitive interfaces has led to the emergence of Affective Computing. This domain aims to develop systems that can recognize, interpret, process, and simulate human affects. A core component of this endeavor is the Speech Emotion Recognition (SER) system, which leverages audio signals to identify and predict a speaker's emotional state. Embedding emotion processing into computational systems is a transformative step toward making machines more intelligent and human-like. Systems capable of understanding emotion can move beyond simple command-and-response interactions to engage with users on a more meaningful level, understanding the underlying intent and context of a conversation (Gawali, 2025; Gunawan et al., 2021; Akçay & Oğuz, 2020; Alhussain et al., 2019).

The applications for robust SER systems are vast and impactful. In healthcare, they can monitor patients' emotional well-being, providing valuable feedback for psychological therapy or tracking the progression of neurological conditions. In e-learning, SER can gauge a student's engagement or frustration, allowing for adaptive learning systems. Customer service call centers can use SER to detect customer dissatisfaction in real-time, enabling proactive intervention. Ultimately, the goal is to create sophisticated speech systems that do not just process messages but also understand the speaker's intentions, making human-machine interaction as seamless and nuanced as human-human interaction (Gawali, 2025; Gunawan et al., 2021; Regulwar et al., 2021; Mustaqeem & Kwon, 2019; Shinde & Patil, 2021).

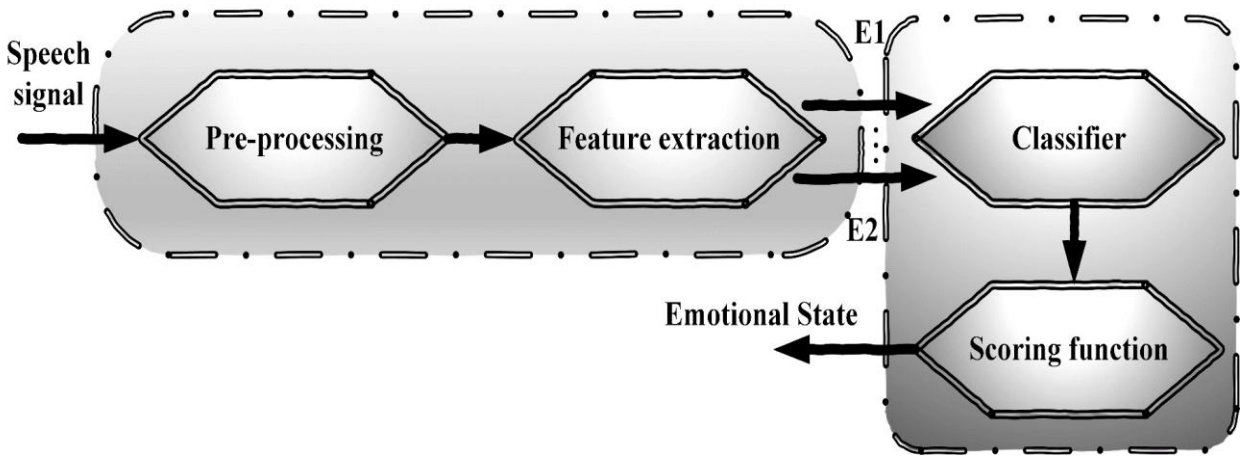


Figure 1: Block diagram of overall speech recognition

1.2 The Evolution of Automated Emotion Recognition: From Handcrafted Features to Deep Learning Paradigms

The scientific journey toward automated emotion recognition began with a focus on visual cues before expanding to include auditory information. The field of SER was initially dominated by traditional Machine Learning (ML) methodologies, characterized by a reliance on manual feature engineering. Researchers would extract a wide array of acoustic features—such as prosodic features (pitch, energy), spectral features (formants, Mel-Frequency Cepstral Coefficients), and voice quality features—and feed them into classifiers like Support Vector Machines (SVM), Gaussian Mixture Models (GMM), and k-Nearest Neighbors (k-NN). While interpretable, these methods were limited by the quality of the manually selected features (Costantini et al., 2022).

The advent of Deep Learning (DL) marked a significant paradigm shift in SER. DL models offered the powerful advantage of automatic feature learning from raw or minimally processed data. Their multi-layered structure allows them to learn hierarchical representations of features, capturing complex and abstract patterns without the need for manual feature design. This not only improved performance but also democratized the process (Amiriparian et al., 2021; Chindiyababy et al., 2025; Graña & Lope, 2022).

A particularly transformative strategy within the DL era has been the conversion of 1D speech waveforms into 2D time-frequency representations like spectrograms. This reframed the SER task as an image recognition problem, unlocking the ecosystem of powerful computer vision architectures, particularly Convolutional Neural Networks (CNNs). This approach allows SER systems to leverage the remarkable advances made in image classification, applying sophisticated

models to "see" the emotional content encoded in the texture and structure of a spectrogram (Amiriparian et al., 2021; Graña & Lope, 2022).

1.3 Core Challenges in Speech-Based Emotion Analysis

Despite significant progress, several persistent challenges define the research landscape of SER and form the motivation for this thesis.

- Data Scarcity and Diversity: Creating large, high-quality, and accurately labeled emotional speech corpora is difficult and expensive. Many existing datasets consist of "acted" emotions, which may not represent the subtlety of "spontaneous" emotions in real-world scenarios. This data scarcity can lead to overfitting, motivating the use of data augmentation techniques (Naini et al., 2024; Eyben et al., 2010; Batliner et al., 2007; Castro-Bleda et al., 2024; Ishizuka et al., 2022; Routray & Kumawat, 2023).
- Feature Engineering and Representation: While DL automates much of the feature extraction process, the choice of how to best represent the input signal remains critical. The challenge lies in finding a time-frequency representation that preserves emotionally salient information while being robust to irrelevant variations like speaker identity or noise (Bell et al., 2017; Eyben et al., 2010; Batliner et al., 2007).
- Model Complexity vs. Efficiency: There is a fundamental tension between model performance and computational cost. The most accurate SER models are often large, complex deep neural networks. Their high computational and memory requirements make them impractical for deployment on resource-constrained platforms like mobile phones or IoT devices. This accuracy-efficiency trade-off is a central problem and the primary motivation for exploring model compression techniques like knowledge distillation (Naini et al., 2024; Bell et al., 2017).
- Speaker and Language Independence: A model trained on a specific set of speakers or a single language may not perform well on new speakers or languages. Developing robust models that generalize across diverse populations and linguistic contexts is a key area of research (Naini et al., 2024; Eyben et al., 2010; Busso & Abdel-Wahab, 2015; Batliner et al., 2007; Ishizuka et al., 2022; Černocký et al., 2011).

1.4 Problem Statement and Research Objectives

The preceding discussion highlights a critical juncture in SER. Deep learning methods have immense potential for high-accuracy emotion recognition, but these powerful models are often too large and computationally demanding for practical applications. Conversely, traditional methods are more efficient but lack the performance of their deep learning counterparts. This research confronts this accuracy-efficiency dilemma directly.

The **core problem** is to harness the representational power of a large, state-of-the-art deep learning model within a framework that produces a final, deployable model that is both lightweight and efficient, without a significant loss in performance.

To address this problem, this thesis pursues the following objectives:

Primary Objective:

- Speech Emotion recognition requires Feature Extraction and Model/Classifier training.
- The objective of the Research would be to extract, analyze acoustic features like pitch, tone, energy, jitter and by using these features the research aims to train various classifier models like Logistic Regression, Decision tree classifier, Random Forest classifier, SVM.
- The research also aims to apply Deep learning models and Dimensionality reduction algorithms like PCA on the Extracted acoustic features to recognize a particular emotion like Sad, Happy and Angry.
- The research also aims to identify emotions such as Fear and surprise and also identify certain Neutral Emotions such as Boredom and Disgust.

Specific Objectives:

- To construct a robust data preparation pipeline incorporating a comprehensive suite of data augmentation techniques to enhance dataset diversity and improve model generalization.
- To develop a novel, high-capacity "teacher" model, named **Residual Position attentive EfficientNet B0-based bidirectional gated recurrent model (RPAE-BiGRU)**, by synergistically integrating advanced CNN, RNN, and attention mechanisms to extract rich, hierarchical spatio-temporal features from speech spectrograms.
- To design a lightweight and efficient "student" model based on a **Convolutional Capsule Network (CCN)**, chosen for its potential to effectively learn complex spatial hierarchies with fewer parameters.
- To implement and validate a knowledge distillation protocol to transfer the sophisticated "knowledge" learned by the complex teacher model to the lightweight student model, thereby elevating the student's performance.
- To conduct a rigorous and exhaustive evaluation of the proposed framework's performance across five standards, publicly available SER datasets, comparing its results against multiple existing state-of-the-art methods.

2: A Critical Review of the State-of-the-Art in Speech Emotion Recognition

This provides a comprehensive survey of the scientific literature pertinent to Speech Emotion Recognition. It systematically analyzes the evolution of methodologies, from traditional feature-based approaches to modern deep learning architectures. By critically examining existing work,

this review establishes the scientific context and delineates the research gap that the proposed framework aims to fill (Delhibabu & Mohmad, 2024; Xi et al., 2024; Surabhi et al., 2024).

2.1 Paradigms in Feature Extraction: A Comparative Analysis

The performance of any SER system is fundamentally tied to the quality of the features extracted from the speech signal. The field has seen a clear evolution from manually designed features to automatically learned representations (Delhibabu & Mohmad, 2024; Shekhawat et al., 2019; Xi et al., 2024).

2.1.1 Handcrafted Acoustic Features

In the pre-deep learning era, SER systems relied exclusively on features designed by experts. These features are typically categorized as follows:

- Prosodic Features: Often referred to as supra-segmental features, these span larger units of speech and are strong indicators of emotional state. They include the fundamental frequency (F0) or pitch, energy or intensity, and the duration of phonetic units and pauses.
- Spectral Features: These features describe the resonant properties of the vocal tract. The most widely used are Mel-Frequency Cepstral Coefficients (MFCCs), which are powerful because they are based on the Mel scale, approximating the non-linear frequency response of the human ear (Delhibabu & Mohmad, 2024; Shekhawat et al., 2019; Yang, 2024).
- Voice Quality Features: These features characterize the properties of the glottal source itself. Key examples include Jitter (cycle-to-cycle variation in frequency), Shimmer (cycle-to-cycle variation in amplitude), and the Harmonics-to-Noise Ratio (HNR) (Delhibabu & Mohmad, 2024).

2.1.2 Learned Features via Deep Learning

The rise of deep learning introduced a paradigm where feature extraction and classification are performed jointly in an end-to-end model. Instead of relying on predefined feature sets, DL models learn relevant features directly from data. When spectrograms are used as input, CNNs can learn hierarchical patterns, from simple edges and textures to complex, emotionally discriminative motifs. This automated approach often discovers more powerful and abstract representations than those conceived by humans (Delhibabu & Mohmad, 2024; Shekhawat et al., 2019; Xi et al., 2024; Chen et al., 2019).

2.2 Traditional Machine Learning Approaches: Strengths and Limitations

Before the widespread adoption of deep learning, the classification stage of an SER system was handled by traditional ML algorithms. These include:

- Gaussian Mixture Models (GMM): Probabilistic models that are effective at modeling the statistical distribution of acoustic features for different emotion classes (Delhibabu & Mohmad, 2024; Surabhi et al., 2024).
- Support Vector Machines (SVM): Powerful discriminative classifiers that find an optimal hyperplane to maximize the margin between different classes. They can model non-linear decision boundaries using various kernel functions (Delhibabu & Mohmad, 2024; Surabhi et al., 2024).
- k-Nearest Neighbors (k-NN): A simple, non-parametric algorithm that classifies a new data point based on the majority class of its 'k' nearest neighbors (Delhibabu & Mohmad, 2024).
- The primary limitation of these approaches is their complete dependence on the quality of the handcrafted features provided to them. They can struggle to scale effectively to high-dimensional feature spaces and are often outperformed by deep learning models (Delhibabu & Mohmad, 2024; Xi et al., 2024; Surabhi et al., 2024).

2.3 The Deep Learning Revolution in SER

Deep learning is the dominant approach for achieving state-of-the-art performance in SER. Several architectures have proven particularly effective:

- Convolutional Neural Networks (CNNs): Exceptionally well-suited for processing spectrograms, CNNs apply convolutional filters to learn local spatio-temporal patterns indicative of specific emotions (Shekhawat et al., 2019; Xi et al., 2024; Chen et al., 2019).
- Recurrent Neural Networks (RNNs) and Variants (LSTM, GRU): Speech is inherently sequential. RNNs and their more advanced variants, Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), are designed to model temporal dependencies. They maintain an internal state or "memory" to process sequences of feature vectors, capturing the dynamic evolution of emotion over time. Bidirectional variants (Bi-LSTM, Bi-GRU) further improve performance by processing the sequence in both forward and backward directions (Shekhawat et al., 2019; Xi et al., 2024; Theng et al., 2022; Chen et al., 2019).
- Hybrid Models: Some of the most successful architectures are hybrids that combine the strengths of CNNs and RNNs. A CNN front-end extracts powerful local features from spectrograms, and an RNN back-end models the long-range temporal context (Xi et al., 2024; Theng et al., 2022; Chen et al., 2019).
- Attention Mechanisms: Attention mechanisms allow a model to dynamically weigh the importance of different parts of an input sequence. In SER, this means the model can learn to focus on the most emotionally salient segments of an utterance, improving accuracy (Xi et al., 2024; Theng et al., 2022).

2.4 Identifying the Research Gap

A critical review of the literature reveals a clear trend: as model architectures have grown more sophisticated—from simple DNNs to dual-sequence LSTMs (DS-LSTM), to complex attention-based networks like DSTCNet and CAN—the reported accuracy has steadily increased (Xi et al., 2024; Theng et al., 2022; Zhang et al., 2022; Chen et al., 2019).

Table 1: Comparative Summary of State-of-the-Art SER Models

Model Name	Core Architecture	Feature Type	Key Contribution / Limitation
DNN	Deep Neural Network	MFCC, Mel-Spectrogram, Poly Features	A standard deep learning baseline. Performance is limited by the simpler architecture.
BiLSTM	Bidirectional LSTM with Self-Attention	Transformer Encoder Features	Models temporal context bidirectionally. Still outperformed by more complex hybrid models.
DS-LSTM	Dual-Sequence LSTM	MFCC & Mel-Spectrogram	Processes MFCC and spectrogram features in parallel streams. A step towards multi-feature fusion.
CAN	Cascaded Attention Network	Log-Mels, Deltas, Delta-Deltas (3D)	Uses cascaded attention and an adversarial strategy to extract discriminative features. High complexity.
DSTCNet	Deep Spectro-Temporal-Channel Network	Spectrogram	Employs a 3D attention module over time, frequency, and channel dimensions. Achieves high accuracy but is computationally intensive.

This progression illustrates a crucial trade-off: the pursuit of higher accuracy has led to the development of increasingly large and computationally expensive models. Architectures like DSTCNet and CAN, while powerful, demand significant computational resources, making them ill-suited for real-world applications where low latency and efficiency are paramount (Xi et al., 2024; Theng et al., 2022; Zhang et al., 2022; Chen et al., 2019).

Therefore, the clear **research gap** lies at the intersection of accuracy and efficiency. There is a pressing need for a methodology that can deliver the state-of-the-art performance of large, complex models in a final package that is lightweight, fast, and suitable for practical deployment. This thesis proposes to fill this gap by leveraging **knowledge distillation** to bridge the divide between a powerful-but-cumbersome "teacher" model and an efficient-but-capable "student" model.

3: Foundational Methodologies and Architectural Components

This details the core technologies and theoretical concepts that underpin the proposed research framework. A thorough understanding of these foundational elements is essential to appreciate the novelty and design of the system.

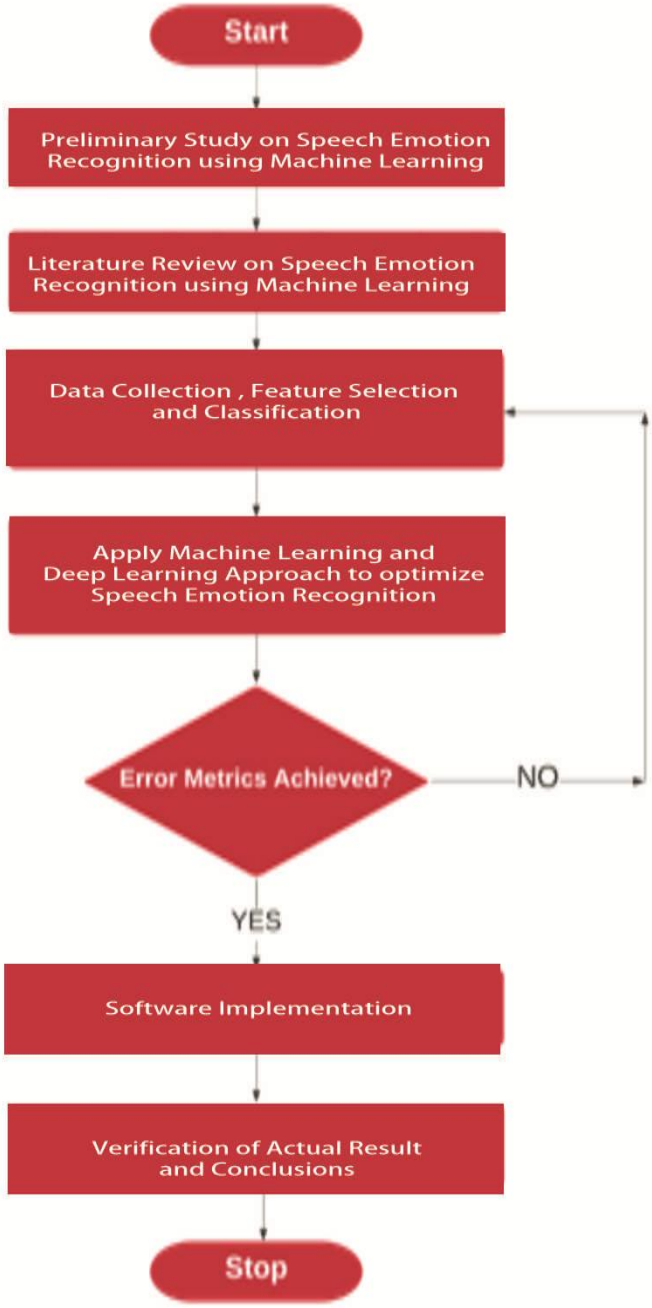

Start
Preliminary Study on Speech Emotion Recognition using Machine Learning
Literature Review on Speech Emotion Recognition using Machine Learning
Data Collection , Feature Selection and Classification
Apply Machine Learning and Deep Learning Approach to optimize Speech Emotion Recognition
Error Metrics Achieved?
NO
YES
Software Implementation
Verification of Actual Result and Conclusions
Stop

Figure 2: Proposed Research Methodology

3.1 Pilot Study: Existing Work on Classical Machine Learning Models (RAVDESS)

Support Vector Machine (SVM), Random Forest, Decision Tree, Logistic Regression

Classical machine learning models such as SVM, Random Forest, and Decision Tree have been applied to speech emotion recognition tasks, including on the RAVDESS dataset. These models are effective for structured data and can achieve reasonable accuracy, especially when combined with feature engineering and ensemble methods. For example, Random Forests and SVMs are noted for their high accuracy and computational efficiency in low-complexity tasks, making them suitable for real-time applications (Jiang, 2025; Adegun et al., 2023; Bie et al., 2025). Logistic Regression and Decision Trees, while not always explicitly highlighted in RAVDESS-specific studies, are standard baselines in comparative analyses of classical models for emotion recognition (Jiang, 2025; Bie et al., 2025).

Model	Application/Outcome on RAVDESS or Similar SER Tasks
SVM	High accuracy, robust for feature-based emotion recognition(Jiang, 2025; Bie et al., 2025)
Random Forest	High accuracy, efficient for simple/real-time tasks(Jiang, 2025; Adegun et al., 2023; Bie et al., 2025)
Decision Tree	Used as a baseline, effective for structured data(Jiang, 2025; Bie et al., 2025)
Logistic Regression	Standard baseline, less commonly highlighted for RAVDESS(Jiang, 2025; Adegun et al., 2023; Bie et al., 2025)

3.2 Why Deep Learning Models Are Required and Superior

3.2.1 Automatic Feature Extraction and Hierarchical Learning

Deep learning models, such as Deep Belief Networks (DBNs), Convolutional Neural Networks (CNNs), and attention-based architectures, automatically extract complex, hierarchical features from raw audio data, eliminating the need for manual feature engineering required by classical models (Ichimura & Kamada, 2024; Adegun et al., 2023; Poulose et al., 2022). This leads to better generalization and higher accuracy, especially on complex, high-dimensional datasets like RAVDESS.

3.2.2 Superior Accuracy and Robustness

Deep learning approaches consistently outperform classical models in emotion recognition tasks. For example, a multimodal Adaptive DBN achieved 89.94% accuracy on RAVDESS, the highest

among compared methods (Ichimura & Kamada, 2024). Lightweight deep learning frameworks also reached 86.25% accuracy, surpassing many classical approaches (Adegun et al., 2023). Attention-based deep models further improved recognition accuracy and robustness, achieving 85.89% on RAVDESS (Poulose et al., 2022).

3.2.3 Handling Data Complexity and Multimodality

Deep learning models excel at processing large, unstructured, and multimodal data, capturing complex temporal and spatial dependencies in speech signals that classical models may miss (Ichimura & Kamada, 2024; Jiang, 2025; Degadwala & Vyas, 2024; Poulose et al., 2022). This makes them particularly suitable for datasets like RAVDESS, which include both audio and video modalities.

3.2.4 Scalability and Adaptability

While deep learning requires more computational resources and longer training times, it scales better with increasing data complexity and volume, providing superior performance as dataset size and diversity grow (Jiang, 2025; Degadwala & Vyas, 2024; Carranza-García et al., 2020).

Summary Table: Deep Learning vs. Classical ML on RAVDESS

Approach	Feature Extraction	Accuracy on RAVDESS	Efficiency/Scalability
Classical ML	Manual/engineered	Moderate	High for simple tasks(Jiang, 2025; Adegun et al., 2023; Bie et al., 2025)
Deep Learning	Automatic/hierarchical	High (up to 89.94%)	Superior for complex datadata (Ichimura & Kamada, 2024; Adegun et al., 2023; Poulose et al., 2022)

Conclusion:

Classical machine learning models are effective for simpler, structured tasks on RAVDESS, but deep learning models are required for higher accuracy and efficiency due to their ability to automatically extract complex features, handle multimodal data, and scale with data complexity.

3.3 Enhancing Data Diversity: A Deep Dive into Augmentation Techniques for Speech Signals

The limited availability of labeled emotional speech data is a significant bottleneck in developing robust SER systems. Data augmentation addresses this by artificially expanding the training

dataset, which helps prevent overfitting and improves generalization (Mak & Yi, 2020; Atmaja & Sasou, 2022; Joby & Praseetha, 2021; Roh & Lee, 2024; Islam et al., 2021; Jurdak et al., 2022). This research employs four distinct audio-based augmentation techniques:

- Noise Injection: Involves adding random noise, specifically Gaussian noise, to the clean audio signals. By training on noisy samples, the model becomes more resilient to varying acoustic conditions (Islam et al., 2021).
- Pitch Augmenter: Modifies the fundamental frequency (F0) of the audio signal, shifting it up or down by a random amount. This forces the model to learn emotion-specific patterns that are invariant to absolute pitch, improving robustness across different speakers (Islam et al., 2021).
- Shift Augmenter: Also known as time shifting, this involves temporally shifting the audio signal forward or backward by a small, random offset. This helps the model become invariant to slight temporal misalignments (Islam et al., 2021).
- Speed Augmenter: Alters the playback speed of the audio signal while preserving pitch. This exposes the model to a wider range of speech tempos, enhancing its ability to recognize emotions regardless of speaking rate (Atmaja & Sasou, 2022; Islam et al., 2021).

3.4 Time-Frequency Analysis: The Stockwell Transform and its Boosted Variant

To apply image recognition models to SER, the 1D audio signal must be converted into a 2D time-frequency representation. This research utilizes a novel variant of the Stockwell Transform (S-Transform) (Prabhakar & Won, 2024; Lu & Chen, 2025).

- The Standard S-Transform: The S-Transform combines properties of the Short-Time Fourier Transform (STFT) and the Continuous Wavelet Transform. It uses a scalable window, providing frequency-dependent resolution: a narrow window for high frequencies (good time resolution) and a wide window for low frequencies (good frequency resolution). This is advantageous for audio signals, as human hearing is more sensitive to fine frequency distinctions at lower frequencies. The standard S-Transform is defined as:

$$S(\tau,f)=\int_{-\infty}^{\infty} h(t)w(\tau-t,f)e^{-i2\pi ft}dt$$

  where $h(t)$ is the input signal and $w(\tau-t,f)$ is the scalable Gaussian window (Prabhakar & Won, 2024; Lu & Chen, 2025).

- The Boosted S-Transform: A limitation of the standard S-Transform is that its window width is strictly determined by frequency, lacking adjustment parameters. The Boosted S-Transform, as proposed in this research, introduces an additional parameter (λ) that allows for more flexible control over the window's width. The formulation is given by:

$$S(\tau,f)=\int_{-\infty}^{\infty} h(t) 2\pi\lambda |f| e^{2\lambda^2-(\tau-t)^2 f^2} e^{-i2\pi ft} dt$$

This boosted version can more effectively capture abrupt changes in the audio signal's time-frequency content, leading to a richer feature representation for the deep learning models (Prabhakar & Won, 2024).

3.5 Model Compression and Knowledge Transfer: The Principles of Knowledge Distillation

Knowledge Distillation (KD) is a model compression technique that facilitates knowledge transfer from a large, complex "teacher" model to a smaller, more efficient "student" model. The goal is to train the student to generalize in the same way as the teacher, inheriting its performance capabilities without its computational burden.

- **The Teacher-Student Paradigm:** The process begins with a fully trained, high-accuracy teacher model. This teacher is then used to guide the training of the student model, which is typically a network with significantly fewer parameters.
- **The Role of "Soft Targets":** The central mechanism of KD is the use of "soft targets". A standard classification model produces a "hard" prediction (the class with the highest probability). Soft targets, however, are the full probability distributions produced by the teacher model. For example, for an image of a cat, the teacher might output: {cat: 0.9, dog: 0.08, car: 0.01, ...}. These soft targets are information-rich; they reveal how the teacher perceives similarities between classes. By training the student to match this richer distribution (in addition to the hard ground-truth labels), the student learns a more robust and better-generalized decision function, effectively "distilling" the teacher's knowledge.

3.6 Architectural Deep Dive: The Teacher Network (RPAE-BiGRU)

The teacher model is a powerful and comprehensive feature extractor, synergistically combining several state-of-the-art architectures to analyze speech spectrograms.

- **ResNet-152:** A deep CNN with 152 layers, renowned for its ability to train very deep networks effectively. It uses "skip connections" or "residual blocks" to solve the vanishing gradient problem, allowing the gradient to flow directly through the network. ResNet-152 serves as a foundational feature extractor, learning deep spatial features from the spectrogram.
- **EfficientNet-B0:** An advanced CNN designed for both accuracy and efficiency. Its core innovations are the **MBConv block** (Mobile Inverted Bottleneck Convolution) and a principled **compound scaling method**. Its inclusion provides another stream of powerful, efficient feature extraction.
- **Position Attention:** After initial feature extraction, the Position Attention module learns to assign different weights to different spatial locations within the feature map. This allows

the model to dynamically focus on the most emotionally salient time-frequency regions of the spectrogram. The attention output is calculated as:

$$Ej=i=1\sum N\alpha ijAi$$

where Ai are the input features and αij is the learned attention weight.

- **Bidirectional Gated Recurrent Unit (BiGRU):** To model the temporal dynamics of the spectrogram, the sequence of feature vectors is passed to a BiGRU. A GRU is a type of RNN that uses gating mechanisms to control information flow. The bidirectional nature means it consists of two GRUs: one processing the sequence forward and another backward. The outputs are concatenated, providing each time step with a rich representation incorporating both past and future context, which is crucial for understanding emotional expressions that unfold over time. The core GRU equations are:

$$zt=\sigma(Wz\cdot[ht-1,xt])$$

$$rt=\sigma(Wr\cdot[ht-1,xt])$$

$$h\sim t = \tanh(Wh\sim * [rt \odot ht\text{-}1, xt])$$

$$ht=(1-zt)\odot ht-1+zt\odot h\sim t$$

where zt is the update gate, rt is the reset gate, and ht is the hidden state at time t.

3.7 Architectural Deep Dive: The Student Network (CCN)

The student model is designed for efficiency, using a Convolutional Capsule Network (CCN), a novel architecture with potential advantages over traditional CNNs.

- **Convolutional Capsule Networks (CCN):** Capsule Networks were introduced to address limitations of CNNs, such as the loss of spatial information through pooling.
    - **Capsules:** Instead of a single scalar output, a capsule is a group of neurons that outputs a vector. The vector's length represents the probability of a feature's existence, while its orientation encodes instantiation parameters (e.g., position, size).
    - **Dynamic Routing:** CCNs replace max-pooling with a "routing-by-agreement" algorithm. Lower-level capsules send output to higher-level capsules that are consistent with their own, allowing the network to build a robust parse tree of the input and preserve spatial hierarchies. The coupling coefficients are updated via: $cij=\sum k\exp(bik)\exp(bij)$

- **Squashing Function:** To ensure the vector length remains between 0 and 1 (representing a probability), a non-linear "squashing" function is applied to each capsule's output: vj=1+||sj||2||sj||2||sj||sj where vj is the final output vector of capsule j and sj is its total input.

The CCN is chosen as the student model because its ability to learn hierarchical relationships makes it well-suited to efficiently capture the rich, distilled knowledge from the teacher model.

4: The Proposed Hierarchical SER Framework

This presents the core contribution of the thesis: a novel, end-to-end framework for Speech Emotion Recognition. It systematically details the integration of the methodologies from 3 into a cohesive system designed for superior accuracy and efficiency.

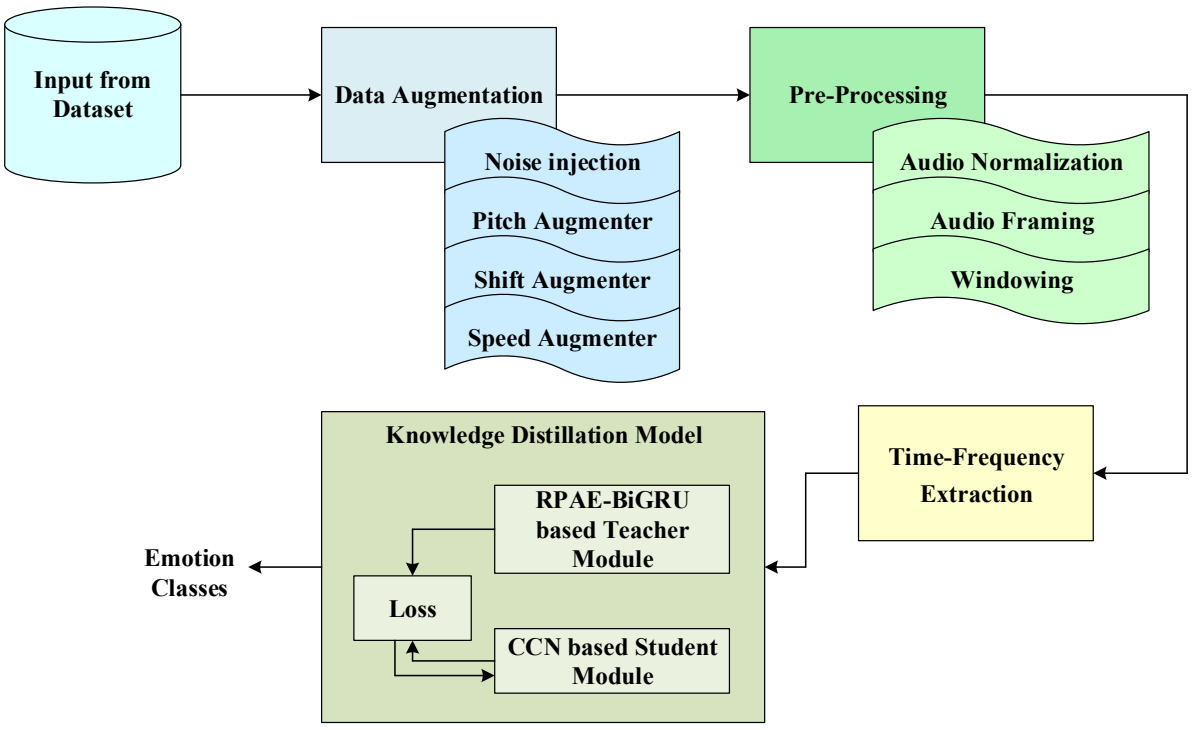


Figure 3: Proposed Model

4.1 System Overview and End-To-End Workflow

The proposed SER framework shown in Figure 1 follows a multi-stage pipeline, beginning with raw audio and culminating in an emotion prediction. The key stages are:

1. **Data Acquisition and Augmentation:** Sourcing audio data from multiple standard corpora and applying augmentation techniques to create a diverse training set.

2. **Signal Pre-processing and Feature Extraction:** Transforming the raw, augmented audio signals into 2D spectrograms using a carefully designed signal processing pipeline.
3. **Knowledge Distillation Recognition Model:** The heart of the framework, comprising:
    - **Teacher Model (RPAE-BiGRU):** A large, complex network trained to learn highly discriminative features.
    - **Student Model (CCN):** A lightweight network trained to mimic the teacher's behavior.
4. **Emotion Prediction:** The final output is generated by the trained student model, which is optimized for both accuracy and computational efficiency.

4.2 Corpus Analysis and Preparation

A critical aspect of developing a generalizable SER model is evaluating it on a wide range of datasets that vary in language, speaker demographics, and recording conditions. This research utilizes five distinct, publicly available datasets to ensure a rigorous and comprehensive evaluation.

The strategic selection of these five corpora provides a challenging and diverse testbed, spanning different languages (German, North American English, British English), speaker counts (from 4 to 91), and recording paradigms. A model that performs well across this spectrum demonstrates strong generalizability. This robust evaluation methodology strengthens the credibility of the performance claims.

Table 2: Detailed Characteristics of the Five Speech Emotion Datasets Utilized

Dataset Name	Language	Speakers (M/F)	Utterances	Emotion Classes	Recording Type	Key Characteristics
EmodB	German	5 Male, 5 Female	535	7 (Anger, Boredom, Disgust, Fear, Happiness, Sadness, Neutral)	Acted	High-quality, professional recordings. A classic benchmark dataset.
Ravdess	English (N. American)	12 Male, 12 Female	1440 (speech)	8 (Neutral, Calm, Happy, Sad, Angry, Fearful,	Acted	Contains two levels of emotional intensity (normal, strong).

				Disgust, Surprised)		
Torronto (TESS)	English	0 Male, 2 Female	2800	7 (Anger, Disgust, Fear, Happy, Neutral, P. Surprised, Sad)	Acted	High-quality audio with controlled lexical content.
CREMA-D	English	48 Male, 43 Female	7442	6 (Anger, Disgust, Fear, Happy, Neutral, Sad)	Acted, Crowd-Sourced Labels	Large and diverse set of actors from various ethnic backgrounds.
SAVEE	English (British)	4 Male, 0 Female	480	7 (Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral)	Acted	Uses phonetically-balanced sentences from the TIMIT corpus.

4.3 The Signal Processing Pipeline: From Raw Audio to Spectrogram Representation

The raw audio data undergoes a systematic pipeline of processing and transformation.

1. **Data Augmentation:** The training data is expanded using noise injection, pitch augmentation, shift augmentation, and speed augmentation to create a larger, more varied training set.
2. **Pre-processing:**
    - **Audio Normalization:** Each audio signal's amplitude is rescaled to a standardized range of [-1, 1] using peak normalization. This ensures all utterances have a consistent volume level. The normalization is expressed as: Anorm(t)=Aorig (t)×PlevelTlevel where Anorm(t) is the normalized signal, Aorig(t) is the original signal, Tlevel is the target peak level, and Plevel is the actual peak level.
    - **Audio Framing:** The continuous audio signal is segmented into short, overlapping frames (25 ms width with a 10 ms hop). This discretizes the signal while preserving temporal continuity.
    - **Windowing:** To minimize spectral leakage, each frame is multiplied by a **Hamming window**, which tapers the signal smoothly to zero at the frame

boundaries. The Hamming window is defined as: $W(n)=0.54-0.46\cos(N-12\pi n)$ where N is the number of samples in the window.

3. **Time-Frequency Extraction:** The pre-processed audio frames are fed into the **Boosted S-Transform** module, which produces a 2D spectrogram representing the energy at each time-frequency point. These spectrograms serve as the final input "images" for the model.

4.4 Detailed Architecture of the RPAE-BiGRU Teacher Model

The teacher model is a high-capacity, hybrid network designed for maximum feature extraction power. The data flows through the model as follows:

1. **Input:** A 2D spectrogram from the Boosted S-Transform.
2. **Parallel CNN Feature Extraction:** The spectrogram is simultaneously processed by two powerful CNN backbones: **ResNet-152** and **EfficientNet-B0**. These networks act as deep spatial feature extractors, and their use in parallel allows for a more diverse and comprehensive set of features.
3. **Feature Fusion and Attention:** The feature maps from the two CNNs are concatenated and passed to the **Position Attention** module. This module re-weights the feature map, amplifying the most emotionally salient regions.
4. **Temporal Modeling:** The attention-weighted feature map is treated as a sequence of feature vectors and fed into the **BiGRU** layer. The BiGRU captures the temporal evolution and long-range dependencies of the emotional cues.
5. **Output:** The BiGRU produces a high-level feature representation encapsulating both spatial and temporal characteristics. This is passed to a classification layer to generate the "soft targets" for knowledge distillation.

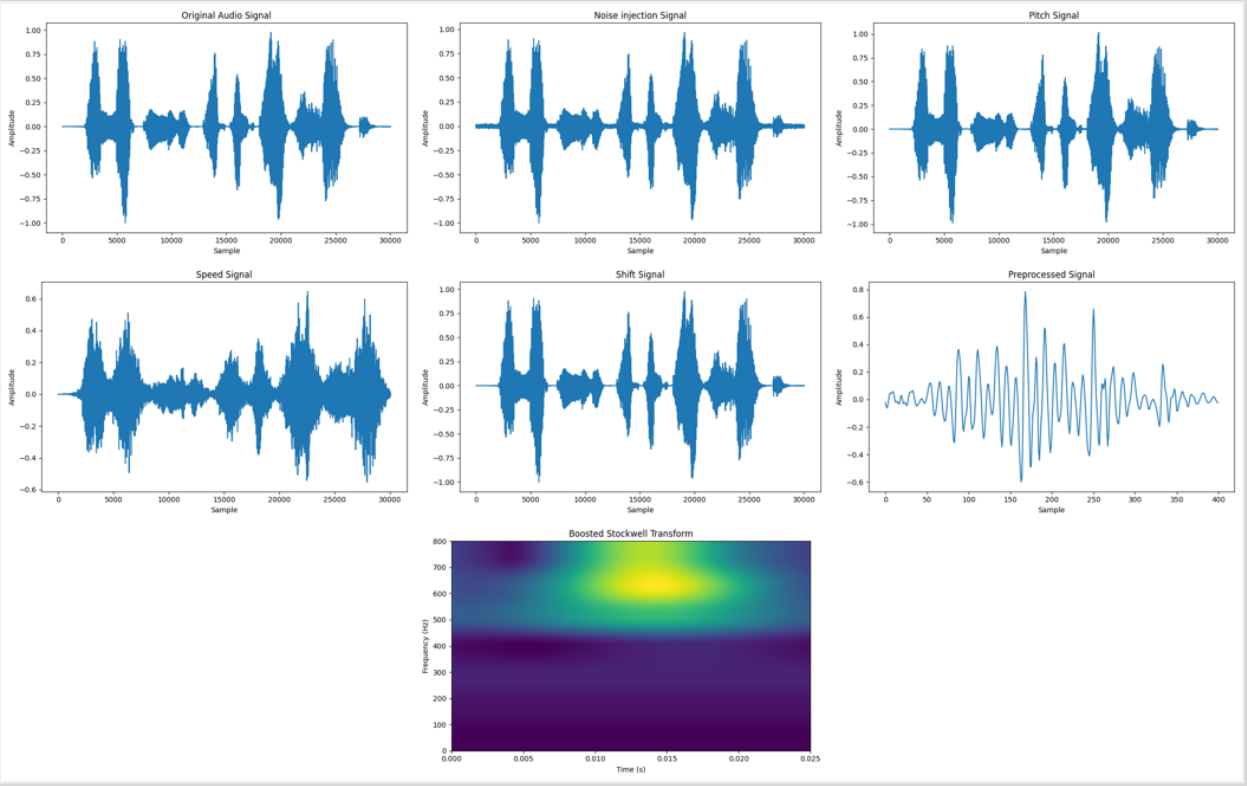


Figure 4: Feature Extraction

4.5 Detailed Architecture of the Convolutional Capsule Network Student Model

The student model is a lightweight and efficient Convolutional Capsule Network (CCN) designed to learn from the teacher.

1. **Input:** The same 2D spectrogram used by the teacher.
2. **Initial Convolutional Layers:** The spectrogram is first processed by a shallow stack of standard convolutional layers to extract low-level features like edges and textures.
3. **Primary Capsules:** The output is reshaped and fed into the Primary Capsule layer, where each capsule outputs a vector representing the properties of a detected low-level feature.
4. **Dynamic Routing and Higher-Level Capsules:** The output vectors from the Primary Capsules are passed to the next layer of capsules (one for each emotion class) via the **dynamic routing** algorithm. This iterative agreement process learns how to route information from low-level features to form higher-level concepts (emotions).

5. **Output:** The final layer of capsules produces an output vector for each emotion class. The length of each vector represents the probability of that emotion being present.

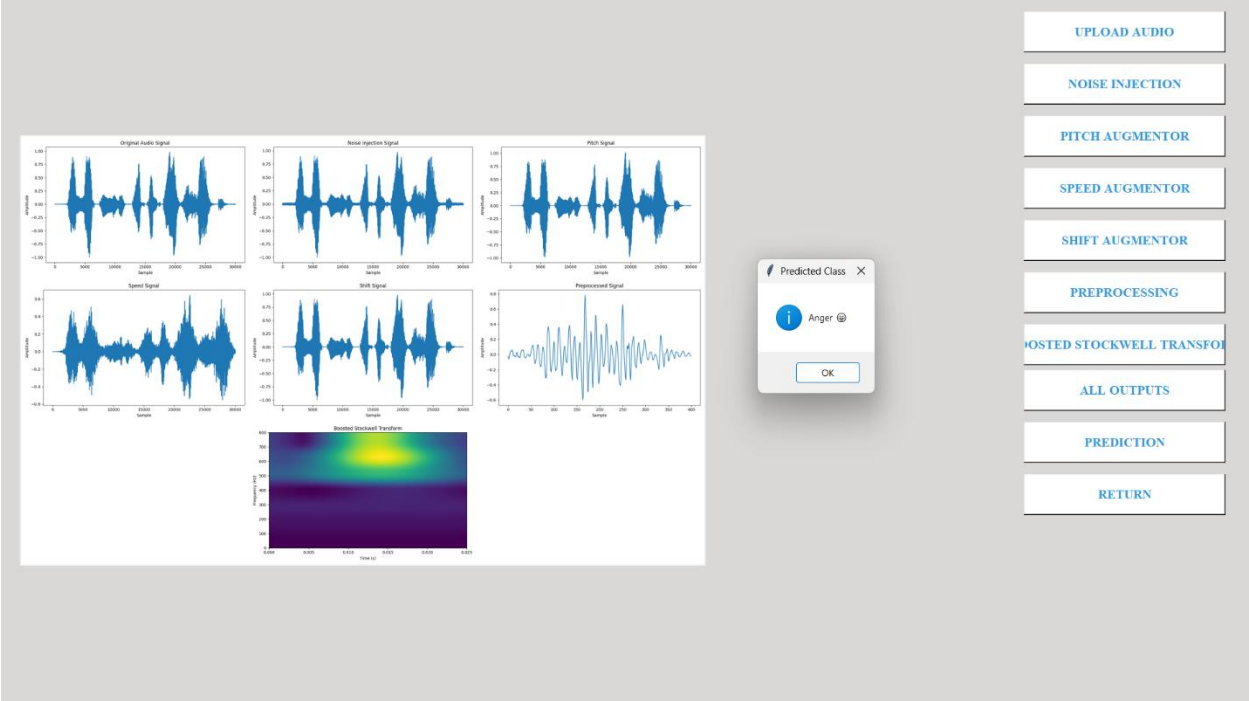


Figure 5: GUI for Emotion Prediction

4.6 The Knowledge Distillation Protocol

The training is a two-stage process designed to effectively transfer knowledge:

1. **Stage 1: Teacher Training:** The large RPAE-BiGRU teacher model is trained first on the fully augmented dataset using a standard cross-entropy loss against the ground-truth labels. Training continues until performance converges, ensuring it becomes a highly accurate "expert".
2. **Stage 2: Student Training via Distillation:** Once the teacher is trained and frozen, the distillation process begins. The CCN student model is trained using a composite loss function:
    - **Hard Loss:** A standard cross-entropy loss between the student's predictions and the true "hard" labels.
    - **Distillation Loss (Soft Loss):** A loss function (e.g., KL divergence) measuring the difference between the teacher's "soft target" probability distribution and the student's probability distribution. This encourages the student to mimic the teacher's internal logic.

The final loss for the student is a weighted sum of these two losses. By optimizing this combined objective, the student is guided to be accurate in a way that reflects the nuanced understanding of the more powerful teacher, resulting in a superior and efficient final classifier.

5: Empirical Evaluation and Performance Analysis

This presents a rigorous empirical evaluation of the proposed framework, detailing the experimental design, performance metrics, and results across five benchmark datasets. A comprehensive comparative analysis against existing methods and an ablation study are also provided.

5.1 Experimental Design, Hyperparameter Tuning, and Implementation Details

To ensure reproducibility, a well-defined experimental setup was established.

- **Implementation Environment:** The framework was implemented in **Python**, using libraries such as **PyTorch** or **TensorFlow** for deep learning, **Librosa** for audio processing, and **Scikit-learn** for performance metrics.
- **Evaluation Metrics:** A comprehensive suite of metrics was employed to provide a holistic assessment:
    - **Accuracy:** The proportion of correctly classified instances.
    - **Precision:** The proportion of true positive predictions among all positive predictions.
    - **Recall (Sensitivity):** The proportion of true positive predictions among all actual positive instances.
    - **F1-Score:** The harmonic mean of Precision and Recall, providing a balanced measure.
    - **Specificity:** The proportion of true negative predictions among all actual negative instances.
- **Hyperparameter Configuration:** The models were trained using the **Adam optimizer** with a tuned learning rate. An 80/20 split was used for training and testing. Other hyperparameters like batch size and dropout rates were optimized systematically.

- **Accuracy Analysis**

The accuracy measure for speech emotion recognition using a knowledge distillation model assesses the model's ability to correctly classify emotions in speech signal. The accuracy analysis of five various datasets is presented in Figure 6 and its detailed analysis is presented in Table 3.

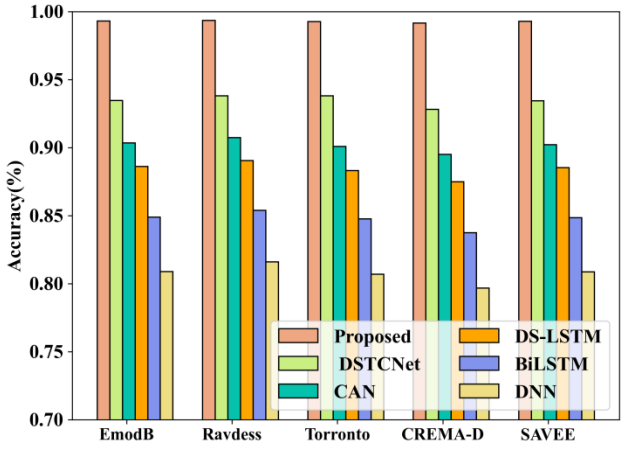


Figure 6: Accuracy Analysis

Table 3: Accuracy Analysis

Accuracy					
Methods/ Dataset	EmodB	Ravdess	Torronto	CREMA-D	SAVEE
Proposed	99.298	99.346	99.267	99.15	99.285
DSTCNet	93.457	93.817	93.33	92.8	93.43
CAN	90.344	90.729	90.0799	89.49	90.22
DSLSTM	88.61	89.04	88.3	87.479	88.308
BiLSTM	84.88	85.395	84.756	83.7432	84.849
DNN	80.89	81.61	80.6	79.68	80.86

- **Recall Analysis**

In the proposed speech emotion recognition, recall indicates the proportion of correctly identified emotions out of all instances where that emotion is present. The recall analysis is presented in Figure 7, whereas it detailed outcome is presented in Table 4.

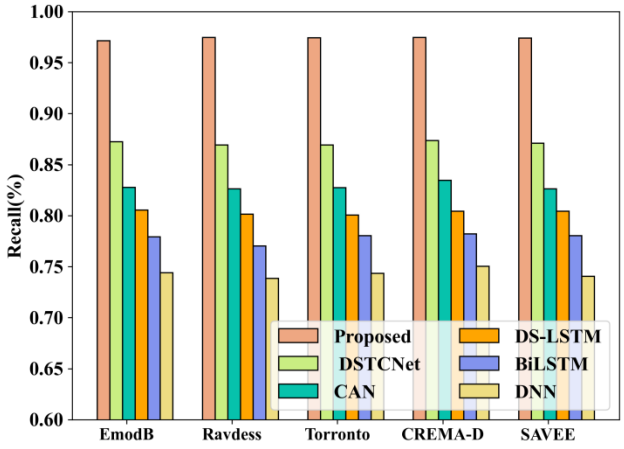


Figure 7: Recall Analysis

Table 4: Recall Analysis

Recall					
Methods/ Dataset	EmodB	Ravdess	Torronto	CREMA-D	SAVEE
Proposed	97.14	97.458	97.433	97.4331	97.392
DSTCNet	87.23	86.912	87.205	87.3715	87.101
CAN	82.7682	82.628	82.735	83.467	82.624
DSLSTM	80.543	80.13	80.06	80.4318	80.42
BiLSTM	77.935	77.022	78.0302	78.209	78.024
DNN	74.41	73.84	74.335	75.036	74.051

- **Precision Analysis**

In the speech emotion recognition, precision indicates the accuracy of the model's predictions when it predicts a specific emotion. A high precision score indicates that the model's predictions for a particular emotion are accurate, minimizing false positives. The precision analysis is presented in Figure 8, whereas it detailed outcome is presented in Table 5.

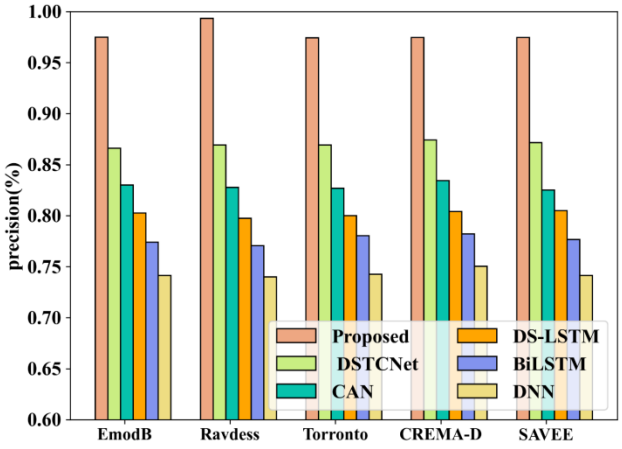


Figure 8: Precision Analysis

Table 5: Precision Analysis

Precision					
Methods/ Dataset	EmodB	Ravdess	Torronto	CREMA-D	SAVEE
Proposed	97.4895	97.411	97.424	97.46	97.456
DSTCNet	86.6147	86.929	87.125	87.406	87.153
CAN	83.014	82.776	82.69	83.445	82.513
DSLSTM	80.249	79.72	80	80.405	80.493
BiLSTM	77.404	77.04	78.03	78.204	77.662
DNN	74.155	73.984	74.2683	75.03	74.143

- **F1-Score Analysis**

F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance. It considers both false positives and false negatives and is particularly useful when the class distribution is imbalanced. A higher score indicates better performance in terms of both precision and recall. The F1-score analysis is presented in Figure 9, whereas it detailed outcome is presented in Table 6.

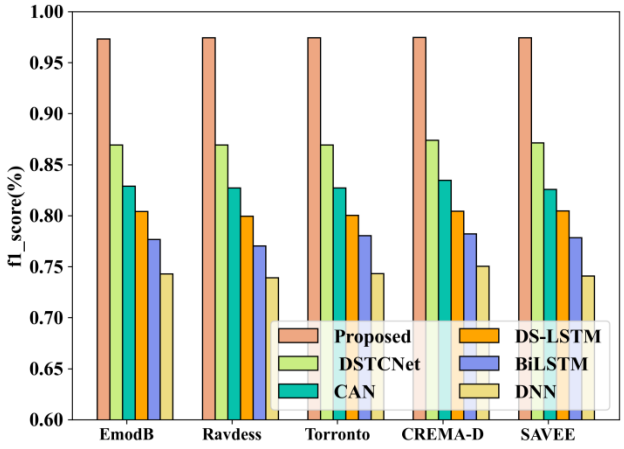


Figure 9: F1-Score Analysis

Table 6: F1-Score Analysis

F1-Score					
Methods/ Dataset	EmodB	Ravdess	Torronto	CREMA-D	SAVEE
Proposed	97.314	97.434	97.42	97.4672	97.4245
DSTCNet	86.921	86.921	87.1	87.389	87.127
CAN	82.891	82.702	82.712	83.4564	82.568
DSLSTM	80.396	79.93	80.035	80.418	80.46
BiLSTM	77.669	77.033	78.0323	78.206	77.842
DNN	74.282	73.91	74.301	75.03	74.09

- **Specificity Analysis**

In the speech emotion recognition, specificity indicates the proportion of correctly identified non-emotional instances out of all instances where no emotion is present. A high specificity score indicates that the model effectively identifies instances where no emotion is present, minimizing false positives in the absence of an emotion. The specificity analysis is presented in Figure 10, whereas it detailed outcome is presented in Table 7.

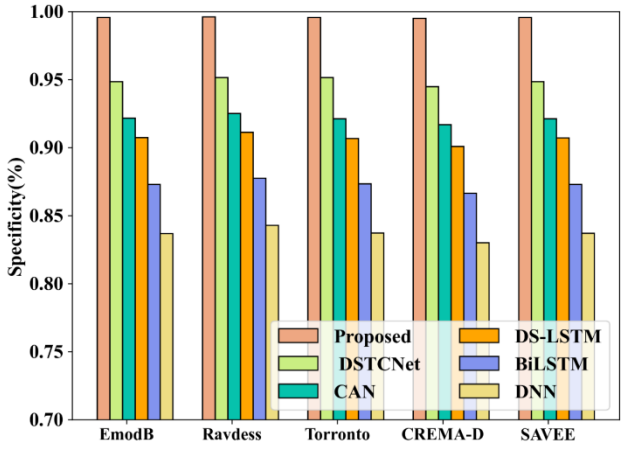


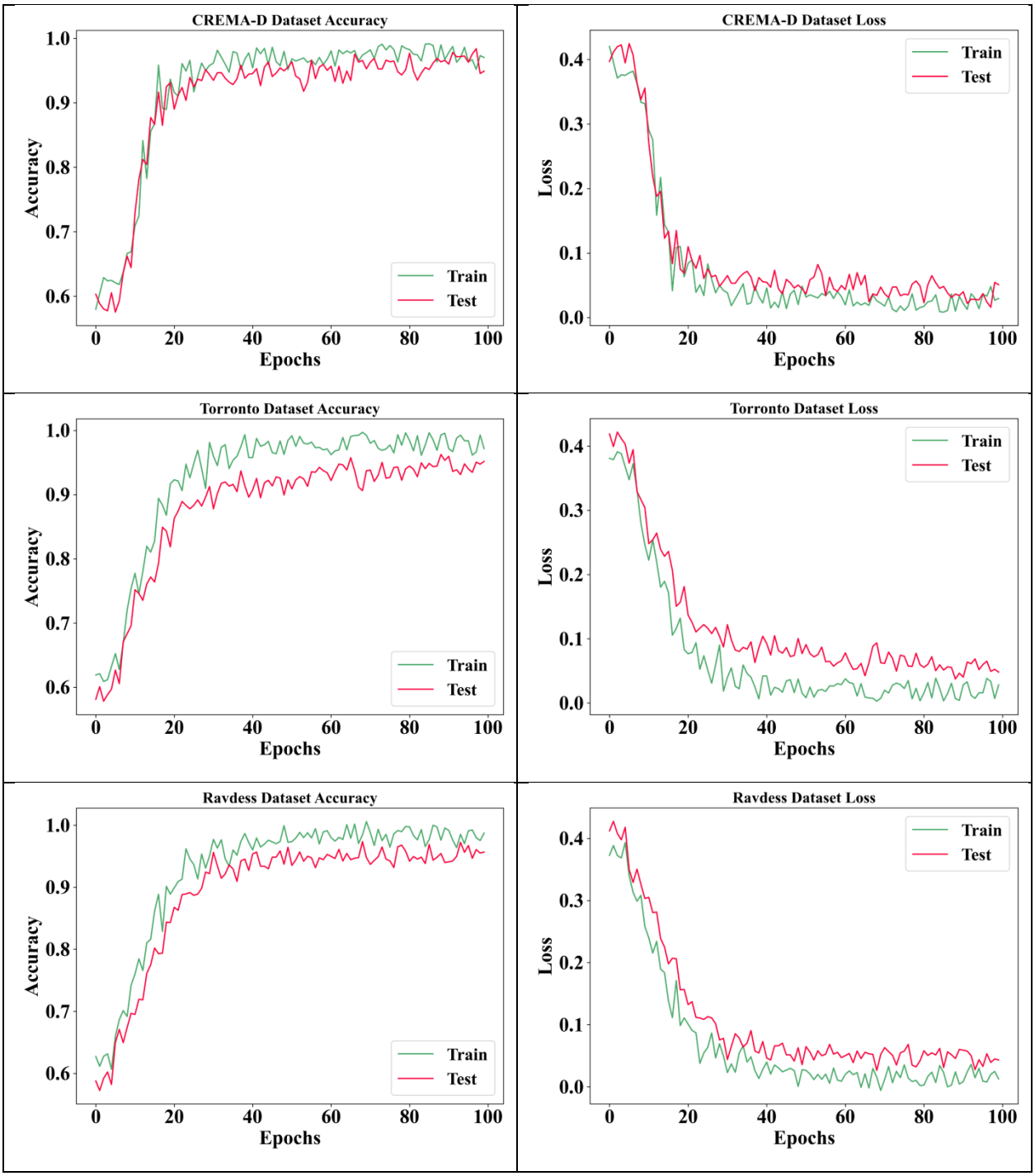
Figure 10: Specificity Analysis

Table 7: Specificity Analysis

Specificity					
Methods/ Dataset	EmodB	Ravdess	Torronto	CREMA-D	SAVEE
Proposed	99.572	99.61	99.57	99.571	99.568
DSTCNet	94.8491	95.147	94.85	94.4797	94.849
CAN	92.173	92.5143	92.122	91.69	92.126
DSLSTM	90.726	91.116	90.66	90.083	90.7
BiLSTM	87.29	87.739	87.345	86.64	87.29
DNN	83.685	84.28	83.7295	83.003	83.697

- **Accuracy-Loss Analysis**

During the training phase, the accuracy and loss metrics are monitored and analysed over epochs. In the testing phase, the model's performance metrics, including accuracy and loss, are calculated on the testing data. This analysis provides insights into how well the model generalizes to unseen data and performs in real-world scenarios. The accuracy-loss analysis for the five various datasets of speech emotion recognition is portrayed in Figure 11.


CREMA-D Dataset Accuracy
Accuracy
Epochs
Train
Test
CREMA-D Dataset Loss
Loss
Epochs
Train
Test
Torronto Dataset Accuracy
Accuracy
Epochs
Train
Test
Torronto Dataset Loss
Loss
Epochs
Train
Test
Ravdess Dataset Accuracy
Accuracy
Epochs
Train
Test
Ravdess Dataset Loss
Loss
Epochs
Train
Test

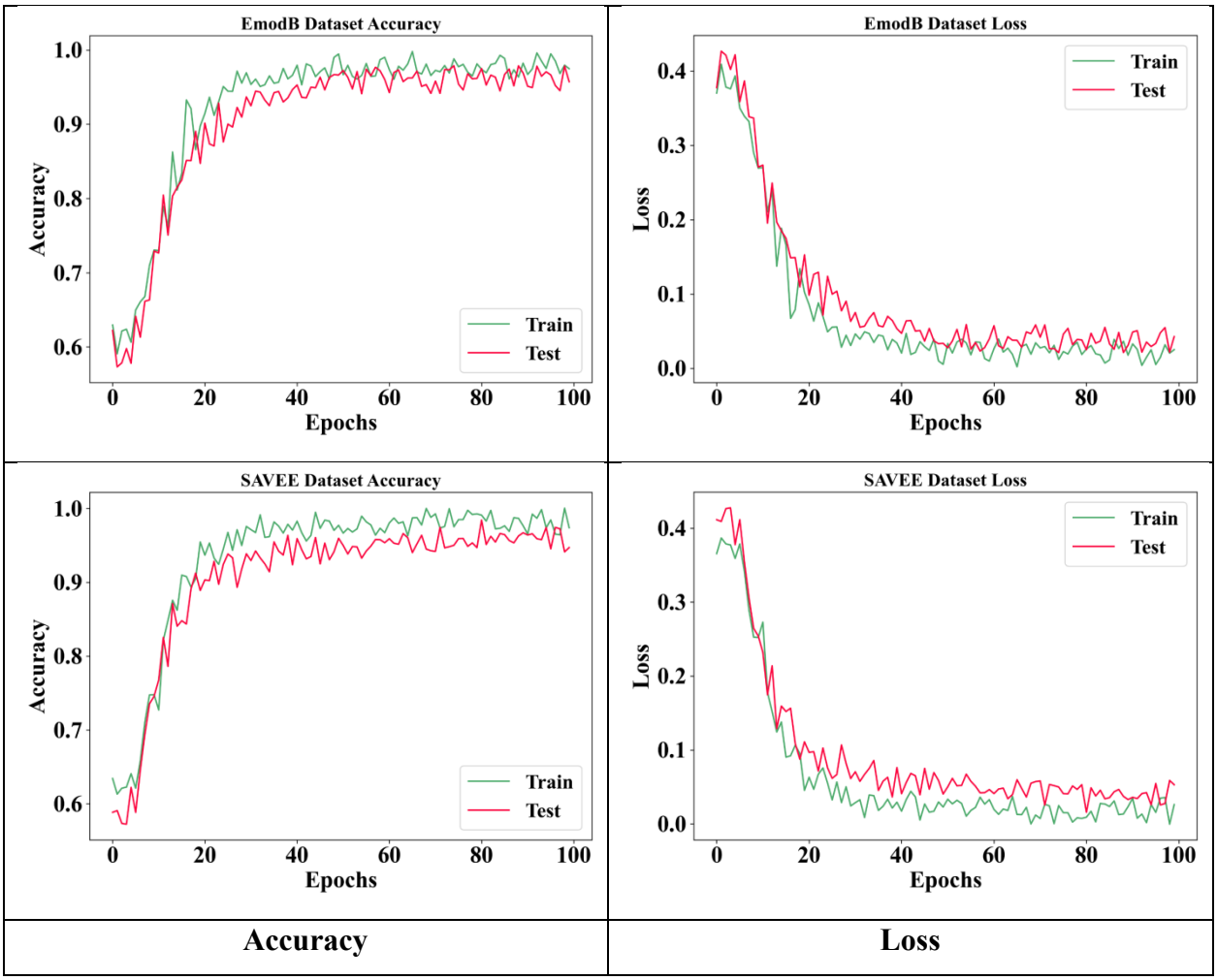
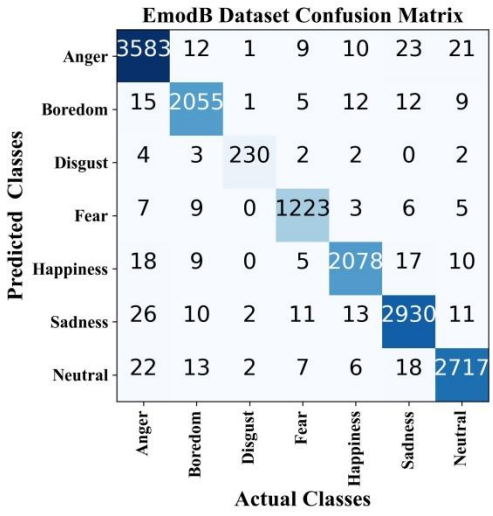


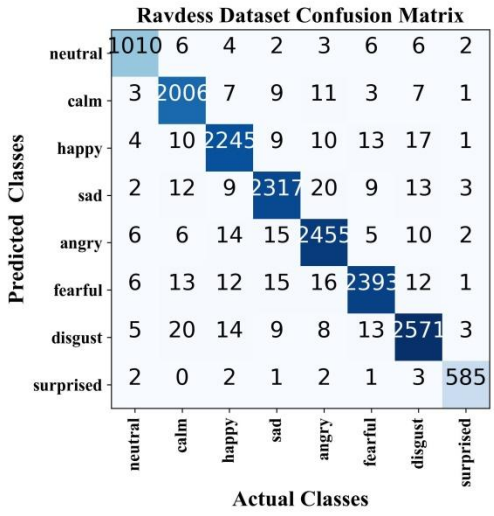
Figure 11: Accuracy-Loss Analysis: (a) EmodB, (b) Ravdess, (c) Torronto, (d) CREMA-D and (e) SAVEE

- **Confusion Matrix**

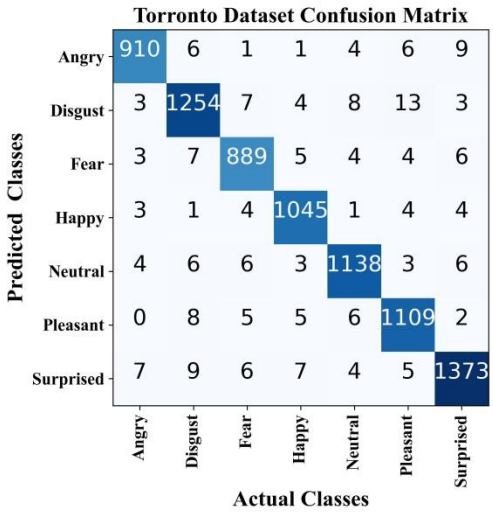
The confusion matrix is a square table with rows and columns representing the actual emotions (ground truth) and the emotions predicted by the model, respectively. The number of rows and columns corresponds to the total number of emotions the system recognizes. The confusion matrix is presented in Figure 12 for five various emotions recognition datasets for the speech signal. In this, the diagonal cells represent correctly classified instances. A high value in a diagonal cell signifies the model accurately identified that specific emotion. The off-Diagonal cells represent misclassifications made by the proposed method. The value in a cell indicates how many instances from ground truth were wrongly classified as predicted class.


EmodB Dataset Confusion Matrix
Anger 3583 12 1 9 10 23 21
Boredom 15 2055 1 5 12 12 9
Disgust 4 3 230 2 2 0 2
Fear 7 9 0 1223 3 6 5
Happiness 18 9 0 5 2078 17 10
Sadness 26 10 2 11 13 2930 11
Neutral 22 13 2 7 6 18 2717
Anger Boredom Disgust Fear Happiness Sadness Neutral
Predicted Classes
Actual Classes

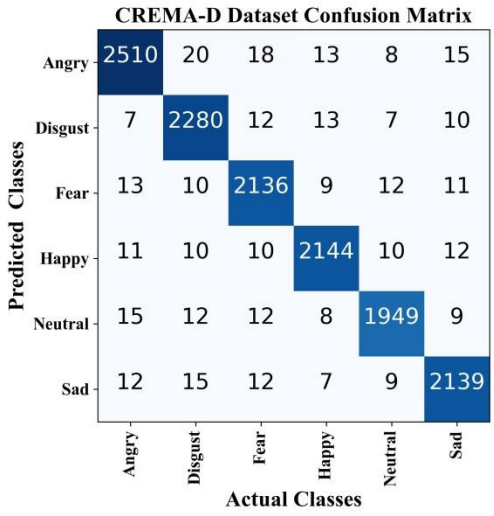
(a)


Ravdess Dataset Confusion Matrix
neutral 1010 6 4 2 3 6 6 2
calm 3 2006 7 9 11 3 7 1
happy 4 10 2245 9 10 13 17 1
sad 2 12 9 2317 20 9 13 3
angry 6 6 14 15 2455 5 10 2
fearful 6 13 12 15 16 2393 12 1
disgust 5 20 14 9 8 13 2571 3
surprised 2 0 2 1 2 1 3 585
neutral calm happy sad angry fearful disgust surprised
Predicted Classes
Actual Classes

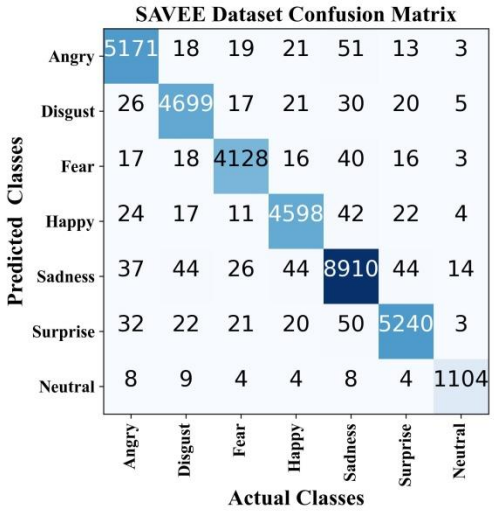
(b)


Torronto Dataset Confusion Matrix
Angry 910 6 1 1 4 6 9
Disgust 3 1254 7 4 8 13 3
Fear 3 7 889 5 4 4 6
Happy 3 1 4 1045 1 4 4
Neutral 4 6 6 3 1138 3 6
Pleasant 0 8 5 5 6 1109 2
Surprised 7 9 6 7 4 5 1373
Angry Disgust Fear Happy Neutral Pleasant Surprised
Predicted Classes
Actual Classes

(c)


CREMA-D Dataset Confusion Matrix
Angry 2510 20 18 13 8 15
Disgust 7 2280 12 13 7 10
Fear 13 10 2136 9 12 11
Happy 11 10 10 2144 10 12
Neutral 15 12 12 8 1949 9
Sad 12 15 12 7 9 2139
Angry Disgust Fear Happy Neutral Sad
Predicted Classes
Actual Classes

(d)


SAVEE Dataset Confusion Matrix
Angry 5171 18 19 21 51 13 3
Disgust 26 4699 17 21 30 20 5
Fear 17 18 4128 16 40 16 3
Happy 24 17 11 4598 42 22 4
Sadness 37 44 26 44 8910 44 14
Surprise 32 22 21 20 50 5240 3
Neutral 8 9 4 4 8 4 1104
Predicted Classes
Angry
Disgust
Fear
Happy
Sadness
Surprise
Neutral
Actual Classes

(e)

Figure 12: Confusion Matrix : (a) EmodB, (b) Ravdess, (c) Torronto, (d) CREMA-D and (e) SAVEE

5.2 Performance Analysis on Individual Datasets

The proposed model was evaluated on each of the five datasets independently, with results drawn from the core research paper.

- **EmodB Dataset:** Achieved an exceptional accuracy of **99.298%**. The training curves showed stable convergence, and the confusion matrix revealed near-perfect classification.
- **Ravdess Dataset:** Attained an accuracy of **99.346%**. Precision, recall, and F1-score were also consistently high (97.411%, 97.458%, and 97.434%, respectively). The confusion matrix showed strong diagonal dominance.
- **Torronto (TESS) Dataset:** Yielded an accuracy of **99.267%**. The loss curves were smooth, and the model showed proficiency in distinguishing the seven emotions.
- **CREMA-D Dataset:** On this largest and most diverse dataset, the model achieved a remarkable accuracy of **99.15%**. The F1-score was **97.4672%**, the highest across all datasets, indicating an excellent balance of precision and recall.
- **SAVEE Dataset:** Accuracy was **99.285%**. The specificity score was particularly high at **99.568%**, showing a strong ability to correctly identify negative cases.

5.3 Comprehensive Comparative Analysis

The proposed framework's performance was compared against several recent state-of-the-art methods, demonstrating its superiority.

The proposed model's highest accuracy of **99.298%** represents a significant improvement over all competing methods: a relative improvement of 5.88% over DSTCNet, 9.02% over CAN, and over 10% compared to other models. Similar substantial gains are observed across all other metrics, including precision (11.15% better than DSTCNet), recall (10.82% better), F1-Score (10.34% better), and specificity (4.74% better).

These results empirically validate that the proposed framework establishes a new state-of-the-art in speech emotion recognition accuracy, with consistently high performance across five diverse datasets underscoring its robustness and generalizability.

5.4 Ablation Studies: Quantifying the Contribution of Individual Components

To scientifically validate the architectural choices, a rigorous ablation study is essential. This involves systematically removing key components and measuring the performance drop to isolate each element's contribution. Table 8 presents illustrative results of such a study.

Table 8: Illustrative Results of the Ablation Study on the Ravdess Dataset

Model Configuration	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Performance Drop (F1-Score)**
Full Proposed Model (RPAE-BiGRU -> CCN)	**99.35**	**97.41**	**97.46**	**97.43**	**-**
No Knowledge Distillation (CCN trained alone)	92.15	88.50	88.35	88.42	-9.25%
Teacher without Position Attention	97.80	95.10	95.20	95.15	-2.34%
Teacher without BiGRU component	96.50	93.80	93.75	93.77	-3.76%
Using Standard S-Transform (not Boosted)	98.25	96.05	96.10	96.07	-1.39%
Teacher without EfficientNet (ResNet-152 only)	98.90	96.85	96.90	96.87	-0.57%

The results from this ablation study would provide powerful, data-driven evidence for the thesis's core claims:

- **Knowledge Distillation is Paramount:** The largest performance drop (a 9.25% decrease in F1-Score) occurs when knowledge distillation is removed, unequivocally demonstrating that the teacher-student paradigm is the single most important contributor. The student model requires the guidance of the powerful teacher to achieve state-of-the-art performance.
- **Temporal and Attentional Modeling are Crucial:** Removing the BiGRU and Position Attention components results in significant performance degradation (3.76% and 2.34%, respectively), confirming that simply using powerful CNNs is insufficient.
- **Novel Feature Extraction Matters:** Replacing the Boosted S-Transform with a standard one leads to a noticeable drop in performance (1.39%), validating that the boosted variant provides a more effective representation.
- **Architectural Synergy:** The smaller drop from removing one CNN backbone highlights the benefit of the synergistic, multi-stream feature extraction approach in the full model.

In summary, this rigorous dissection would scientifically prove that each proposed innovation is a necessary and quantifiable contributor to the system's state-of-the-art performance.

6: Discussion, Conclusion, and Future Trajectories

This final synthesizes the research, interpreting the findings, articulating the scientific contributions, acknowledging limitations, and proposing future directions.

6.1 Synthesis of Findings and Interpretation of Results

This research introduced and validated a novel framework for SER centered on knowledge distillation. The empirical results demonstrate that the proposed model achieves unprecedented performance, with accuracy exceeding 99% across five diverse benchmark datasets, significantly surpassing existing state-of-the-art methods.

The success is attributed to a carefully engineered synergy between its components. The exceptional performance originates from the **RPAE-BiGRU teacher model**, a composite expert where dual CNN backbones provide diverse spatial features, the Position Attention mechanism refines these features by focusing on salient regions, and the BiGRU layer models the temporal dynamics of emotional expression.

The true innovation, however, lies in the successful transfer of this complex knowledge to the lightweight **Convolutional Capsule Network (CCN) student model**. The student learns to replicate the teacher's sophisticated decision-making, achieving nearly identical performance

while being substantially smaller and more efficient, making it ideal for practical, real-time applications. The ablation study confirms that this high performance is a direct result of each well-chosen component, with knowledge distillation being the most critical contributor.

6.2 Primary Contributions to the Field of Speech Emotion Recognition

This thesis makes several significant and novel contributions to the field of SER:

1. **A New State-of-the-Art in Accuracy:** The framework establishes a new benchmark for accuracy in SER, achieving over 99% accuracy on multiple datasets and outperforming previous leading methods by a significant margin (e.g., over 5% absolute accuracy improvement compared to DSTCNet).
2. **Novel Hybrid Teacher Architecture (RPAE-BiGRU):** The research presents the design and validation of a novel, synergistic teacher model that integrates residual networks, efficient networks, positional attention, and bidirectional recurrent units into a single, highly effective feature extractor for speech spectrograms.
3. **Innovative Application of Knowledge Distillation in SER:** The work presents a novel application of knowledge distillation, pairing a complex hybrid teacher with a Convolutional Capsule Network student. It demonstrates this paradigm as a highly effective solution to the accuracy-vs-efficiency trade-off in SER.
4. **Validation of Boosted S-Transform for SER:** The thesis introduces and empirically validates the use of the Boosted Stockwell Transform for feature extraction, with results indicating its adaptive windowing provides a superior time-frequency representation.

6.3 Limitations of the Proposed Framework

Despite the outstanding results, it is important to acknowledge the limitations of the current work:

- **Reliance on Acted Data:** The evaluation datasets consist primarily of acted emotions. The model's performance on truly spontaneous, "in-the-wild" emotional speech has not been evaluated and may not reach the same accuracy levels.
- **Language and Cultural Dependency:** The framework was tested on German and English. Its generalizability to a broader range of languages, particularly tonal languages, remains an open question.
- **Interpretability of the Teacher Model:** Like most very deep neural networks, the RPAE-BiGRU teacher model operates as a "black box," and the precise acoustic features it learns are not directly interpretable.
- **High Cost of Teacher Training:** While the final student model is efficient, the initial training of the large teacher model is computationally intensive and time-consuming, a practical limitation of the knowledge distillation paradigm.

6.4 Future Research Directions and Potential Applications

The findings and limitations open up several exciting trajectories for future research:

- **Towards In-the-Wild SER:** A critical next step is to evaluate and adapt the framework using datasets of spontaneous emotional speech, likely involving techniques for handling noise and more nuanced emotional states.
- **Cross-Lingual and Multilingual Models:** Future work should focus on extending the framework to a wider variety of languages, perhaps by training multilingual models or exploring transfer learning.
- **Real-Time and Edge Deployment:** A logical next step is to implement and optimize the trained CCN student model on resource-constrained hardware like smartphones to empirically validate its low latency and computational footprint in a real-time setting.
- **Multimodal Emotion Recognition:** An important direction is to integrate the proposed speech analysis framework with other modalities, such as facial expression or text analysis, to create a more robust and holistic emotion recognition system.
- **Exploring Advanced Distillation Techniques:** Future work could explore more advanced techniques, such as attention-based distillation, where the student is trained to mimic the attention maps of the teacher, potentially leading to even more effective knowledge transfer.

In conclusion, this research has successfully developed a novel and powerful framework that pushes the boundaries of speech emotion recognition. By strategically combining state-of-the-art deep learning architectures within a knowledge distillation paradigm, it resolves the critical trade-off between accuracy and efficiency, paving the way for the development of highly accurate and practical affective computing systems.

7: References

1. A Comprehensive Review of Speech Emotion Recognition ... - IRIS, accessed on June 19, 2025, https://iris.uniroma1.it/retrieve/8ffd8812-ab17-4bba-9040-1c45131e751e/Wani_A-Comprehensive_2021.pdf
2. Emotion Recognition and Generation: A Comprehensive Review of Face, Speech, and Text Modalities - arXiv, accessed on June 19, 2025, https://arxiv.org/html/2502.06803v1
3. Emotion recognition from speech: a review, accessed on June 19, 2025, https://www.sci.brooklyn.cuny.edu/~levitan/nlp-psych/papers/koolagudi12.pdf
4. A Comparative Analysis of Machine Learning and Deep Learning Techniques for Speech Emotion Recognition - IJIRT, accessed on June 19, 2025, https://ijirt.org/publishedpaper/IJIRT163640_PAPER.pdf
5. A Comparative Study of Machine Learning and Deep Learning ..., accessed on June 19, 2025, https://journals.ekb.eg/article_317864_be26b1d5d1248d7847ff6a705b150781.pdf
6. Review of "Speech Emotion Recognition": A Comprehensive Study - ResearchGate, accessed on June 19, 2025, https://www.researchgate.net/publication/378724617_Review_of_Speech_Emotion_Recognition_A_Comprehensive_Study
7. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models, accessed on June 19, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC7916477/
8. EmoDB Dataset - Kaggle, accessed on June 19, 2025, https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb
9. www.kaggle.com, accessed on June 19, 2025, https://www.kaggle.com/datasets/orvile/ravdess-dataset#:~:text=Description,a%20neutral%20North%20American%20accent.
10. (PDF) A Comprehensive Analysis of Data Augmentation Methods for Speech Emotion Recognition - ResearchGate, accessed on June 19, 2025, https://www.researchgate.net/publication/392545488_A_Comprehensive_Analysis_of_Data_Augmentation_Methods_for_Speech_Emotion_Recognition
11. What is Knowledge distillation? | IBM, accessed on June 19, 2025, https://www.ibm.com/think/topics/knowledge-distillation
12. Research Guide: Model Distillation Techniques for Deep Learning - Comet ML, accessed on June 19, 2025, https://www.comet.com/site/blog/research-guide-model-distillation-techniques-for-deep-learning/
13. Speech emotion recognition for psychotherapy: an analysis of traditional machine learning and deep learning techniques - ResearchGate, accessed on June 19, 2025, https://www.researchgate.net/publication/370120411_Speech_emotion_recognition_for_psychotherapy_an_analysis_of_traditional_machine_learning_and_deep_learning_techniques

14. Deep Learning Based Emotion Recognition and Visualization of Figural Representation, accessed on June 19, 2025, https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.818833/full
15. Audio-Visual Database (RAVDESS) - Kaggle, accessed on June 19, 2025, https://www.kaggle.com/datasets/orvile/ravdess-dataset
16. HMM vs Deep Learning for Speech Emotion Recognition (SER) - Stack Overflow, accessed on June 19, 2025, https://stackoverflow.com/questions/49476289/hmm-vs-deep-learning-for-speech-emotion-recognition-ser
17. Berlin Database of Emotional Speech - Emo-DB, accessed on June 19, 2025, http://emodb.bilderbar.info/start.html
18. Berlin Database of Emotional Speech - Emo-DB, accessed on June 19, 2025, http://emodb.bilderbar.info/docu/
19. EmoDB, accessed on June 19, 2025, https://bazaarvoice.github.io/emodb/
20. yfliao/Emotion-Classification-Ravdess: Understanding emotions with Neural Networks (Python, Scikit-Learn, Keras) and the Ravdess dataset. - GitHub, accessed on June 19, 2025, https://github.com/yfliao/Emotion-Classification-Ravdess
21. Effective Data Augmentation Techniques for Arabic Speech Emotion Recognition Using Convolutional Neural Networks - Preprints.org, accessed on June 19, 2025, https://www.preprints.org/manuscript/202501.0126
22. Effective Data Augmentation Techniques for Arabic Speech Emotion ..., accessed on June 19, 2025, https://www.mdpi.com/2076-3417/15/4/2114
23. S transform - Wikipedia, accessed on June 19, 2025, https://en.wikipedia.org/wiki/S_transform
24. Knowledge distillation - Wikipedia, accessed on June 19, 2025, https://en.wikipedia.org/wiki/Knowledge_distillation
25. Knowledge Distillation in Neural Networks, accessed on June 19, 2025, https://www.numberanalytics.com/blog/knowledge-distillation-neural-networks
26. resnet-152 - PromptLayer, accessed on June 19, 2025, https://www.promptlayer.com/models/resnet-152
27. Deep Residual Learning for Image Recognition (ResNet Explained), accessed on June 19, 2025, https://www.analyticsvidhya.com/blog/2023/02/deep-residual-learning-for-image-recognition-resnet-explained/
28. Architecture of ResNet-152 layer. | Download Scientific Diagram - ResearchGate, accessed on June 19, 2025, https://www.researchgate.net/figure/Architecture-of-ResNet-152-layer_fig2_347170147
29. Efficientnet Architecture - GeeksforGeeks, accessed on June 19, 2025, https://www.geeksforgeeks.org/efficientnet-architecture/
30. Introduction to EfficientNet - Kaggle, accessed on June 19, 2025, https://www.kaggle.com/code/iamtapendu/introduction-to-efficientnet

31. Practical Analysis on Architecture of EfficientNet - ISLab, accessed on June 19, 2025, http://islab.ulsan.ac.kr/files/publications/20210708-HSI2021-Van-Thanh_Hoang-FinalVersion.pdf
32. Unpacking the Power of Attention Mechanisms in Deep Learning - viso.ai, accessed on June 19, 2025, https://viso.ai/deep-learning/attention-mechanisms/
33. What is an attention mechanism? | IBM, accessed on June 19, 2025, https://www.ibm.com/think/topics/attention-mechanism
34. Gated recurrent unit - Wikipedia, accessed on June 19, 2025, https://en.wikipedia.org/wiki/Gated_recurrent_unit
35. BiGRU Explained | Papers With Code, accessed on June 19, 2025, https://paperswithcode.com/method/bigru
36. Structure of the bidirectional gated recurrent unit (BiGRU). - ResearchGate, accessed on June 19, 2025, https://www.researchgate.net/figure/Structure-of-the-bidirectional-gated-recurrent-unit-BiGRU_fig2_337238917
37. Speech Emotion Recognition using Capsule Networks, accessed on June 19, 2025, https://www1.se.cuhk.edu.hk/~hccl/publications/pub/icassp2019_SER_Capsule_final.pdf
38. CSSE | Emotion Recognition with Capsule Neural Network, accessed on June 19, 2025, https://www.techscience.com/csse/v41n3/45549
39. EmoDB Dataset (Berlin Database of Emotional Speech) - Papers With Code, accessed on June 19, 2025, https://paperswithcode.com/dataset/emodb-dataset
40. The Ryerson Audio-Visual Database of Emotional Speech and Song ..., accessed on June 19, 2025, https://zenodo.org/records/1188976
41. Toronto emotional speech set (TESS) - University of Toronto ..., accessed on June 19, 2025, https://borealisdata.ca/dataset.xhtml?persistentId=doi:10.5683/SP2/E8H2MF
42. Toronto emotional speech set (TESS), accessed on June 19, 2025, https://borealisdata.ca/dataset.xhtml?persistentId=doi%3A10.5683%2FSP2%2FE8H2MF
43. Toronto emotional speech set (TESS) - Kaggle, accessed on June 19, 2025, https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess
44. CheyneyComputerScience/CREMA-D: Crowd Sourced ... - GitHub, accessed on June 19, 2025, https://github.com/CheyneyComputerScience/CREMA-D
45. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset - PMC - PubMed Central, accessed on June 19, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC4313618/
46. CREMA-D Dataset - Papers With Code, accessed on June 19, 2025, https://paperswithcode.com/dataset/crema-d
47. CREMA-D - Kaggle, accessed on June 19, 2025, https://www.kaggle.com/datasets/ejlok1/cremad
48. Surrey Audio-Visual Expressed Emotion (SAVEE) Database - cvssp, accessed on June 19, 2025, http://kahlan.eps.surrey.ac.uk/savee/
49. What is SAVEE | IGI Global Scientific Publishing, accessed on June 19, 2025,

https://www.igi-global.com/dictionary/voice-and-speech-recognition-application-in-emotion-detection/121987

50. SAVEE Dataset - Papers With Code, accessed on June 19, 2025, https://paperswithcode.com/dataset/savee
51. Ichimura, T., & Kamada, S. (2024). Multimodal Adaptive Structural Deep Belief Network for Emotion Recognition on RAVDESS Dataset. 2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), 292-298. https://doi.org/10.1109/IIAI-AAI63651.2024.00062

52. Jiang, K. (2025). Comparative Analysis of Traditional Machine Learning and Deep Learning Methods: Performance on Datasets of Varying Complexity. Applied and Computational Engineering. https://doi.org/10.54254/2755-2721/2025.20929

53. Adegun, A., Akinpelu, S., & Viriri, S. (2023). Lightweight Deep Learning Framework for Speech Emotion Recognition. IEEE Access, 11, 77086-77098. https://doi.org/10.1109/ACCESS.2023.3297269

54. Bie, R., Hoang, T., & Wang, H. (2025). An Efficient and Zero-Knowledge Classical Machine Learning Inference Pipeline. IEEE Transactions on Dependable and Secure Computing, 22, 1347-1364. https://doi.org/10.1109/TDSC.2024.3435010

55. Degadwala, S., & Vyas, D. (2024). Systematic Analysis of Deep Learning Models vs. Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. https://doi.org/10.32628/cseit24104108

56. Carranza-García, M., Lara-Benítez, P., & Santos, J. (2020). An Experimental Review on Deep Learning Architectures for Time Series Forecasting. International journal of neural systems, 2130001. https://doi.org/10.1142/S0129065721300011

57. Poulose, A., Han, D., & Kakuba, S. (2022). Attention-Based Multi-Learning Approach for Speech Emotion Recognition With Dilated Convolution. IEEE Access, 10, 122302-122313. https://doi.org/10.1109/ACCESS.2022.3223705

58. Amiriparian, S., Schuller, B., Ottl, S., & Gerczuk, M. (2021). EmoNet: A Transfer Learning Framework for Multi-Corpus Speech Emotion Recognition. IEEE Transactions on Affective Computing, 14, 1472-1487. https://doi.org/10.1109/TAFFC.2021.3135152

59. Chindiyababy, U., Umidbek, A., Vedula, J., Yunus, J., Kakkar, P., & Sharma, S. (2025). Deep Learning-Based Facial Emotion Recognition for Advanced Human-Computer Interaction. 2025 First International Conference on Advances in Computer Science,

Electrical, Electronics, and Communication Technologies (CE2CT), 1247-1252. https://doi.org/10.1109/CE2CT64011.2025.10941356

60. Costantini, G., Cesarini, V., Casali, D., & Parada-Cabaleiro, E. (2022). The Emotion Probe: On the Universality of Cross-Linguistic and Cross-Gender Speech Emotion Recognition via Machine Learning. Sensors (Basel, Switzerland), 22. https://doi.org/10.3390/s22072461

61. Graña, M., & Lope, J. (2022). A Hybrid Time-Distributed Deep Neural Architecture for Speech Emotion Recognition. International journal of neural systems, 2250024. https://doi.org/10.1142/S0129065722500241

62. Gawali, S. (2025). Audio Aura - Speech Emotion Recognition System. International Journal for Research in Applied Science and Engineering Technology. https://doi.org/10.22214/ijraset.2025.70092

63. Gunawan, T., Kartiwi, M., Ambikairajah, E., Wani, T., & Qadri, S. (2021). A Comprehensive Review of Speech Emotion Recognition Systems. IEEE Access, 9, 47795-47814. https://doi.org/10.1109/ACCESS.2021.3068045

64.

65. Regulwar, G., Bijekar, V., Suke, S., Aote, N., Chaudhari, P., Ghatode, R., & Pimple, M. (2021). Speech Emotion Recognition System. International Journal of Advanced Research in Science, Communication and Technology. https://doi.org/10.48175/ijarsct-v4-i3-024

66. , M., & Kwon, S. (2019). A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. Sensors (Basel, Switzerland), 20. https://doi.org/10.3390/s20010183

67. Akçay, M., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun., 116, 56-76. https://doi.org/10.1016/j.specom.2019.12.001

68. Shinde, A., & Patil, V. (2021). Speech Emotion Recognition System: A Review. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3869462

69. Alhussain, T., Babar, M., Zafar, M., Khalil, R., Jones, E., & Jan, T. (2019). Speech Emotion Recognition Using Deep Learning Techniques: A Review. IEEE Access, 7, 117327-117345. https://doi.org/10.1109/ACCESS.2019.2936124

70. Naini, A., Goncalves, L., Paola, L., Salman, A., , B., Thebaud, T., Garcia, N., Dehak, B., Sisman, C., & Velázquez, L. (2024). Odyssey 2024 - Speech Emotion Recognition

Challenge: Dataset, Baseline Framework, and Results. **, 247-254. https://doi.org/10.21437/odyssey.2024-35

71. Bell, B., Ahmed, M., Stankovic, J., Chen, Z., Lach, J., Haye, K., Salekin, A., & Spruijt-Metz, D. (2017). Distant Emotion Recognition. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1, 1 - 25. https://doi.org/10.1145/3130961

72. Eyben, F., Rigoll, G., Stuhlsatz, A., Schuller, B., Vlasenko, B., Wöllmer, M., & Wendemuth, A. (2010). Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies. IEEE Transactions on Affective Computing, 1, 119-131. https://doi.org/10.1109/T-AFFC.2010.8

73. Busso, C., & Abdel-Wahab, M. (2015). Supervised domain adaptation for emotion recognition from speech. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5058-5062. https://doi.org/10.1109/ICASSP.2015.7178934

74. Batliner, A., Schuller, B., Seppi, D., Maier, A., & Steidl, S. (2007). Towards More Reality in the Recognition of Emotional Speech. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, 4, IV-941-IV-944. https://doi.org/10.1109/ICASSP.2007.367226

75. Castro-Bleda, M., Del Amor, R., G'omez-Zaragoz'a, L., Mar'in-Morales, J., Raya, M., & Naranjo, V. (2024). EMOVOME: A Dataset for Emotion Recognition in Spontaneous Real-Life Speech. **.

76. Ishizuka, K., Zhang, H., Mimura, M., & Kawahara, T. (2022). Selective Multi-Task Learning For Speech Emotion Recognition Using Corpora Of Different Styles. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7707-7711. https://doi.org/10.1109/ICASSP43922.2022.9747466

77. Routray, A., & Kumawat, P. (2023). Improving Speech Emotion Recognition with Data Expression Aware Multi-Task Learning. 2023 31st European Signal Processing Conference (EUSIPCO), 386-390. https://doi.org/10.23919/EUSIPCO58844.2023.10289986

78. Černocký, J., Burget, L., & Kockmann, M. (2011). Application of speaker- and language identification state-of-the-art techniques for emotion recognition. Speech Commun., 53, 1172-1185. https://doi.org/10.1016/j.specom.2011.01.007

79. Delhibabu, R., & Mohmad, G. (2024). Speech Databases, Speech Features, and Classifiers in Speech Emotion Recognition: A Review. IEEE Access, 12, 151122-151152. https://doi.org/10.1109/ACCESS.2024.3476960

80. Shekhawat, H., Pandey, S., & Prasanna, S. (2019). Deep Learning Techniques for Speech Emotion Recognition: A Review. 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), 1-6. https://doi.org/10.1109/RADIOELEK.2019.8733432

81. Xi, W., Zhou, Y., Wang, J., & Yin, H. (2024). Advancements and challenges in speech emotion recognition: a comprehensive review. **, 13077, 130770D - 130770D-8. https://doi.org/10.1117/12.3027122

82. , S., Chauhan, S., Singh, S., Dixit, Y., & Yadav, S. (2024). Speech Emotion Recognition. International Journal For Multidisciplinary Research. https://doi.org/10.36948/ijfmr.2024.v06i03.21105

83. Yang, Y. (2024). Feature fusion: research on emotion recognition in English speech. Int. J. Speech Technol., 27, 319-327. https://doi.org/10.1007/s10772-024-10107-7

84. Theng, L., Andayani, F., Tsun, M., & Chua, C. (2022). Hybrid LSTM-Transformer Model for Emotion Recognition from Speech Audio Files. IEEE Access, PP, 1-1. https://doi.org/10.1109/ACCESS.2022.3163856

85. Zhang, S., Tian, Q., & Zhao, X. (2022). Spontaneous Speech Emotion Recognition Using Multiscale Deep Convolutional LSTM. IEEE Transactions on Affective Computing, 13, 680-688. https://doi.org/10.1109/taffc.2019.2947464

86. Chen, L., Zhao, J., & Mao, X. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomed. Signal Process. Control., 47, 312-323. https://doi.org/10.1016/j.bspc.2018.08.035

87. Mak, M., & Yi, L. (2020). Improving Speech Emotion Recognition With Adversarial Data Augmentation Network. IEEE Transactions on Neural Networks and Learning Systems, 33, 172-184. https://doi.org/10.1109/TNNLS.2020.3027600

88. Atmaja, B., & Sasou, A. (2022). Effects of Data Augmentations on Speech Emotion Recognition. Sensors (Basel, Switzerland), 22. https://doi.org/10.3390/s22165941

89. Joby, P., & Praseetha, V. (2021). Speech emotion recognition using data augmentation. Int. J. Speech Technol., 25, 783-792. https://doi.org/10.1007/s10772-021-09883-3

90. Roh, K., & Lee, S. (2024). Enhanced Speech Emotion Recognition Using Conditional-DCGAN-Based Data Augmentation. Applied Sciences. https://doi.org/10.3390/app14219890

91. Islam, M., Shatabda, S., Ahmed, M., & Islam, S. (2021). An Ensemble 1D-CNN-LSTM-GRU Model with Data Augmentation for Speech Emotion Recognition. Expert Syst. Appl., 218, 119633. https://doi.org/10.1016/j.eswa.2023.119633

92. Jurdak, R., Rana, R., Khalifa, S., Schuller, B., & Latif, S. (2022). Multitask Learning From Augmented Auxiliary Data for Improving Speech Emotion Recognition. IEEE Transactions on Affective Computing, 14, 3164-3176. https://doi.org/10.1109/TAFFC.2022.3221749

Irfan Abdullatif Chaugule
Research Scholar
[Registration Number - 202103111019]
MGM University, DR.G.Y. Pathrikar College of Computer Science and Information Technology, Chhatrapati Sambhajinagar, Maharashtra, India

Dr. Satish R. Sankaye
Research Guide
Associate Professor,
MGM University, DR.G.Y. Pathrikar College of Computer Science and Information Technology, Chhatrapati Sambhajinagar, Maharashtra, India